О вероятности переобучения пороговых конъюнкций

Андрей Ивахненко

Вычислительный Центр им. А. А. Дородницына РАН

Интеллектуализация Обработки Информации, ИОИ-8
18–22 октября 2010, Кипр, г. Пафос

Содержание

Задача классификации

- Выборка объектов: $\mathbb{X}^L = (x_i)_{i=1}^L$.
- Объекты описаны n действительными признаками, $x_i = (x_i^1, \dots, x_i^n)$.
- Каждому объекту x_i соответствует ответ y_i из заданного конечного множества Y.
- Алгоритм классификации $a\colon \mathbb{R}^n \to Y$ — взвешенное голосование логических правил (rules):

$$a(x) = \arg\max_{y \in Y} \sum_{r \in R_y} w_r r(x),$$

где w_r — вес правила r, обычно неотрицательный;
R_y — множество правил класса y.

Правила — конъюнкции пороговых предикатов

В общем случае *правило* — это функция вида

$$r\colon \mathbb{R}^n \to \{0,1\} \in R.$$

В данной работе правила — это семейство конъюнкций пороговых предикатов:

$$r(x) \equiv r(x; c^1, \dots, c^n) = \prod_{j \in \omega} \bigl[x^j \lesseqgtr_j c^j\bigr],$$

$x = (x^1, \dots, x^n) \in \mathbb{R}^n$ — произвольный объект;
$\omega \subseteq \{1, \dots, n\}$ — подмножество признаков;
$\lesseqgtr_j$ — одна из операций сравнения $\{\leqslant, \geqslant\}$;
c^j — порог по j-му признаку.

Хорошие правила

Для поиска (*индукции*) правил класса y по обучающей выборке $X \subset \mathbb{X}^L$ решается задача двухкритериальной оптимизации:

$$P(r, X) = \sum_{x_i \in X} r(x_i)\,[y_i = y] \to \max_r;$$

$$N(r, X) = \sum_{x_i \in X} r(x_i)\,[y_i \neq y] \to \min_r;$$

На практике используются различные критерии информативности $I(P, N)$:

- энтропийный критерий;
- индекс Джини;
- статистические тесты (точный тест Фишера, χ^2, ω^2 и др.)

Структура классов эквивалентности правил

Два правила эквивалентны, $r \sim r'$, если их векторы ошибок совпадают.

Можно использовать разные индикаторы ошибок L:

- $r \overset{p}{\sim} r'$ при $L_p(r, x_i) = [r(x_i) = 0][y_i = y]$;
- $r \overset{n}{\sim} r'$ при $L_n(r, x_i) = [r(x_i) = 1][y_i \neq y]$;
- $r \overset{err}{\sim} r'$ при $L(r, x_i) = [r(x_i) \neq [y_i = y]] = L_p(r, x_i) + L_n(r, x_i)$.

Эквивалентность по информативности:

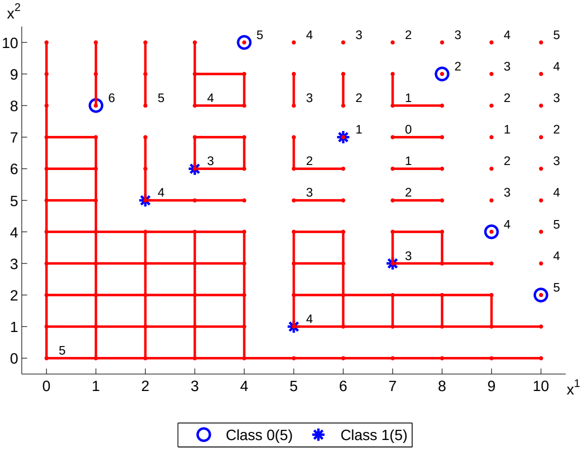
- $r \overset{I}{\sim} r'$ при $I(P, N) = I(P', N')$.

Лемма

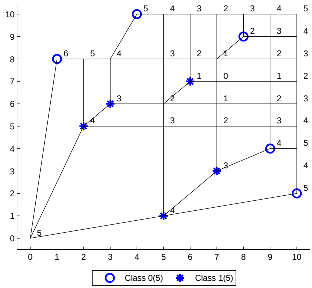
Если правило $r \overset{err}{\sim} r'$, то $r \overset{I}{\sim} r'$, где $I(P, N)$ — функционал информативности.

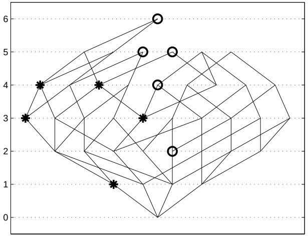
Пример структуры классов эквивалентности $\overset{err}{\sim}$

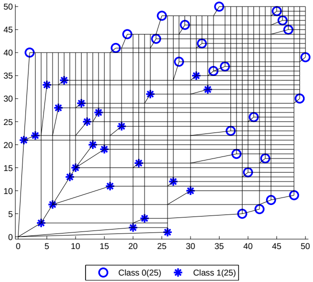
$$r(x) = \prod_{j\in\omega} [x^j \leqslant c^j]$$

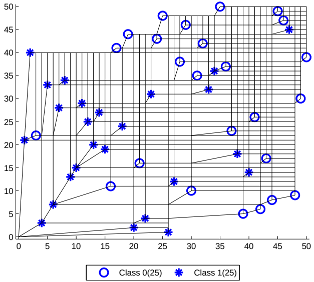


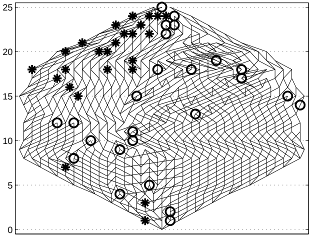
Граф расслоения-связности

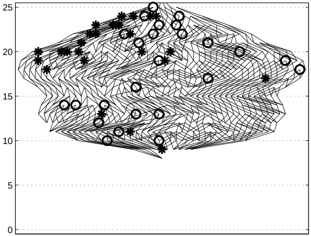





Class 0(25)
Class 1(25)


Class 0(25)
Class 1(25)





Множество объектов $D(q,r) \subseteq \mathbb{X}^L$ *ухудшает* правило q по сравнению с правилом r, если:
1) $r(x_i) = q(x_i)$ для всех $x_i \in \mathbb{X}^L \backslash D(q,r)$;
2) $q(x_i) = [y_i = y]$, $r(x_i) = [y_i \neq y]$ для всех $x_i \in D(q,r)$.

Теорема

Пусть множество объектов $D(q,r)$ ухудшает правило r по сравнению с правилом q. Тогда, чтобы пессимистичный метод МЭР выбрал правило r, объекты множества $D(q,r)$ должны находиться в X'_r.

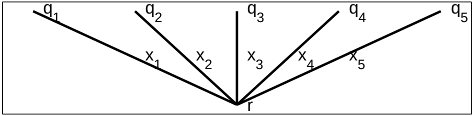
Теорема

Пусть множество объектов $D(q,r)$ состоит из одного объекта x. Тогда, чтобы пессимистичный метод МЭР выбрал правило q, объект должен находиться в X_q.

Метод порождающих и запрещающих множеств

Пример.

$x_i \in D(q_i, r)$



$$[\mu X = r] \leqslant [X_r \subseteq X][X'_r \subseteq \bar{X}]$$

Теорема

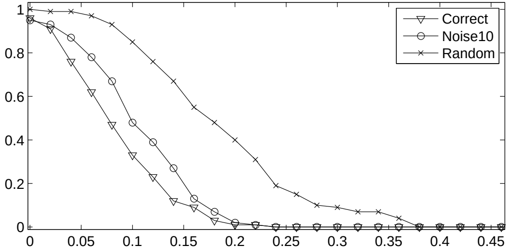
$$Q_\varepsilon(\mu, \mathbb{X}^L) \leqslant \sum_{r \in R} P_r H_{L_r}^{\ell_r, m_r}(s_r(\varepsilon)), \quad P_r = C_{L_r}^{\ell_r} / C_L^{\ell},$$

где $m_r = m(r, \mathbb{X}^L \backslash X_r \backslash X'_r)$,
$L_r = L - |X_r \cup X'_r|,\ \ell_r = \ell - |X_r|$,
$s_r(\varepsilon) = \frac{\ell}{L}(m(r, \mathbb{X}^L) - \varepsilon k) - m(r, X_r)$.

Оценки вероятности переобучения по Монте-Карло

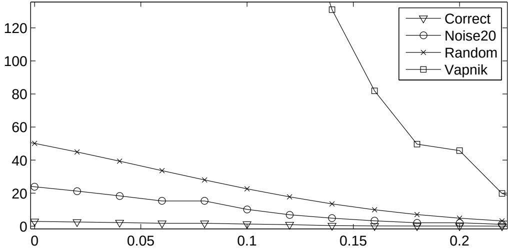
Эксперимент на модельных данных: число признаков $n = 2$, число классов $Y = \{0, 1\}$, число объектов $L = 100$, по 50 объектов в каждом классе. Возьмём три модельных выборки, отличающихся только классификацией объектов.

- Correct— существует правило, разделяющее два класса без ошибок;
- Noise10 — получается из Correct небольшим зашумлением: для 10 пограничных объектов класс меняется на противоположный;
- Random — случайное назначение классов объектам.



Сравнение с оценкой Вапника-Червоненкиса

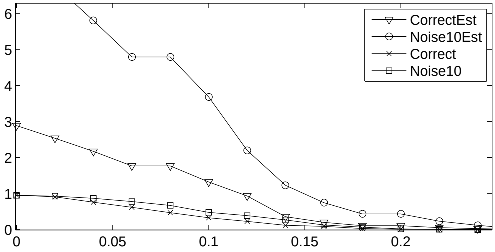
Оценки для выборок Random, Correct, Noise10. Оценка Вапника-Червоненкиса (одинаковая для всех).



Завышенность оценки не велика по сравнению с оценкой Вапника-Червоненкиса.

Сравнение оценок для выборок разной зашумленности

Оценки для выборок Correct и Noise10 по сравнению с точными значениями полученными с помощью метода Монте-Карло.



Завышенность оценки тем больше, чем менее точной является закономерность, содержащаяся в выборке.

Результаты и выводы

- Описана структура классов эквивалентности для пороговых конъюнкций.
- Получены оценки вероятности переобучения, учитывающие расслоение и связность.
- Получен критерий информативности правил, учитывающий возможное переобучение связанное с выбором порогов.